Episode Book

EPISODE #5626

What Multimodal Models Actually Learn From Video

PUBLISHED	28 September 2026 at 16:50 UTC
DURATION	21:31
LISTEN	myweirdprompts.com/episode/multimodal-training-video-data/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.2
BOOK GENERATED	19 October 2026 at 17:38 UTC

SYNOPSIS

Multimodal models are trained on video, audio, and text — but what does "multimodal" actually mean under the hood? This episode digs into late fusion versus early fusion architectures, why video is the largest slice of Emu3.5's 13-trillion-token diet, and a phenomenon called "vision laziness" that explains why rich visual data can sit unused. Then it gets to the question a listener named Daniel asked: if a model learns from video, can you retrieve that knowledge just by typing at it? The answer is yes — but only for some kinds of knowledge, and the asymmetry is stranger than you'd expect.

TRANSCRIPT



Corn

Okay. I need to say something before we start, because it's been building for weeks.



Herman

Go on.



Corn

Every time we do one of these video model episodes, I come away thinking the same thing. The benchmarks say the thing can't do physics. And then I watch the thing do something that looks an awful lot like physics.



Herman

And Daniel noticed.



Corn

Daniel noticed. He wrote in about the Omni episode specifically. His point was that we spent the back half of that episode talking about how the model still fails the strict physics benchmarks, and that the caveat kind of steamrolled the actual news. Because the failure modes we're talking about are crude. People walking through walls. Objects melting into each other. A model that clears that bar is still a different animal from one that doesn't.



Herman

Right. The bar is on the floor and clearing it is still progress.



Corn

But here's where he takes the turn. He says we mentioned the novelty was that Gemini was trained in part on video, and that visual data encodes an understanding that reading text can't. Text is humans describing the world. Video is the world. And that observation cracked something open for him, because he realized he'd never once questioned whether multimodal models are actually trained on multimedia. Or whether multimodal just means text with a vision adapter bolted on the side.



Herman

That's the question.



Corn

So he's got four of them. How have multimodal models traditionally been trained, and what modalities actually go in. Do video generation models have to be trained on video data. Is Omni's novelty the multimedia data, or is it that the segmentation fencing off which format goes into which training pipeline has broken down. And then the big one. If a model is trained on text, audio and video together, does the knowledge encoded in the video bleed into the model in a way a user could retrieve just by typing at it. Text in, text out.



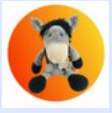
Herman

That last one is the whole episode.



Corn

It's the whole episode. And he wants the under-the-hood version. How media type has traditionally been handled in training pipelines, and what the new ones hint at.



Herman

So we start with what multimodal training has actually meant in practice, because the word has been doing a lot of unearned work.



Corn

Define the term.



Herman

The dominant paradigm for years was late fusion. You take a vision encoder that's already been trained, you take a language model that's already been trained, and you bolt them together with a shallow projector in the middle. LLaVA, DeepSeek-VL, Qwen-Image. The language model stays frozen. The projector learns to translate visual features into something the language model will accept as tokens.



Corn

So the vision is a guest in the language model's house.



Herman

A guest who has to speak the host's language. And that's the critique the Tencent Youtu Lab roadmap makes. Their phrase is that these non-native compositions suffer from a fundamental blindness to raw sensory signals. Because rich visual signals are forced to conform to a pre-existing language space.



Corn

Forced to conform. Meaning the visual information gets squeezed through a channel that was built for words.



Herman

Then you get mid fusion, where features get injected into a joint backbone instead of being stapled on the outside. Qwen3-VL, InternVL-3.5. Better, but still a text-first model that's learned to accommodate images.



Corn

And the third one.



Herman

Early fusion. Born-native. All the modalities go through one unified tokenizer into a single shared embedding space from the very beginning. Transfusion, Chameleon, AnyGPT, Emu3.5. The roadmap writes it as a Transformer over the union of the modality tokenizers. No separate frozen encoders anywhere in the stack.



Corn

So the difference isn't how much data. It's whether there's a wall in the architecture.



Herman

And here's the tension for the episode. The intuitive story about Omni is that the novelty is the data. They trained it on video. The research suggests the novelty is the wall coming down. And that without the wall coming down, the video data may not buy you much at all.



Corn

Say that last part again, because that's the part I want to sit with.



Herman

The Meta paper on the physics of multimodal pretraining is blunt about it. Unifying modalities from the very early stages and training them jointly is more effective than late alignment or sequential training. And they name a phenomenon. Vision laziness. When integration is delayed, the model relies on language priors and under-optimizes the vision components. It learns that it can get most of the way there by guessing from the text, so it stops looking.



Corn

So you can hand a model the entire visual record of human civilization and it'll just... not look at it.



Herman

It'll glance at it and then go back to what it already believed. Which is a very human failure pattern, actually.



Corn

It's the student who reads the summary instead of the book.



Herman

And gets a passing grade, which is the problem. So the segmentation breakdown isn't a nice-to-have. It's the thing that determines whether the video data does any work.



Corn

Alright. Question two. Do video generation models have to be trained on video data.



Herman

In practice, yes. And the strongest evidence isn't an argument, it's a number. Emu3.5, out of BAAI. Pretrained on roughly thirteen trillion tokens, primarily derived from sequential frames and transcripts of internet videos. Sixty-three million videos. Average six and a half minutes. Something like seven hundred and ninety years of continuous footage.



Corn

Seven hundred and ninety years.



Herman

Continuous. If you started watching at the founding of the Ming dynasty and never stopped.



Corn

I'd have finished. I want that on the record. I'd have finished.



Herman

You'd have watched it at your own pace and the sun would have gone out. But here's the part that matters. Their data sampling ratio. Fifty-five percent of pretraining data was video-interleaved. In both stages. Text was eighteen to twenty percent. Image-text pairs sixteen to twenty. Video-text pairs five to eight.



Corn

So video isn't an add-on. It's the largest single slice of the diet.



Herman

By a wide margin. The text is the side dish. And Emu3.5's own framing is that text alone provides only a limited view of the world, that vision is the primary modality through which humans perceive and learn. They explicitly contrast their approach with the conventional one, which relies on paired data made of short, independent samples.



Corn

Short independent samples. Like a caption on a photo.



Herman

A caption on a photo. Versus six and a half minutes of a person doing something in a room, with continuity, with cause and effect, with the camera moving.



Corn

So the answer to Daniel's question is yes, and more than yes. Video models aren't trained on text descriptions of video. They're trained on video.



Herman

Which raises the obvious follow-up. If the data is that rich, why doesn't the model understand physics?



Corn

That's question four and you're getting ahead of me.



Herman

I know. But it's the same question, because the answer to question three is what determines whether the video data turns into knowledge.



Corn

Then answer question three. Is Omni's novelty the data or the architecture.



Herman

The architecture. Google's own model card confirms the data side. Trained on audio, video, image and text, with audio and video annotated with text captions at different levels of detail. But the model card doesn't specify data mixing ratios. It doesn't specify architecture beyond transformer-based with native multimodal support. That's all you get.



Corn

So the disclosure is thin.



Herman

Thin. But Pichai's framing is not thin. He said Gemini was their first model to be natively multimodal, and that training it on a combination of text, code, audio, images and video would give it a deeper understanding of the world. Natively. That's the word doing the work.



Corn

Not trained on multiple modalities. Natively multimodal.



Herman

And the roadmap gives you the taxonomy that makes sense of it. Native models split by input and output. Multi-to-text, which is comprehension, text out. Multi-to-target, generation of one non-text modality. And multi-to-multi, symmetric any-to-any. Omni sits in the last two.



Corn

And the claim is that the early fusion is what makes any-to-any possible at all.



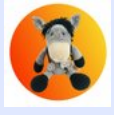
Herman

The paper found specific architectural choices that promote cross-modal synergy. Shared attention and normalization, with modality-specific feed-forward layers. So the attention is shared across modalities, the processing stays partly separate.



Corn

Shared attention, separate feed-forward. That's a specific answer to a question I didn't know had a specific answer.



Herman

And then the counterintuitive finding. Simple tasks act as cross-modal boosters. Low-complexity tasks in one modality improve the other modality. But complex visual distributions, the really rich video stuff, can degrade text perplexity.



Corn

So the easy stuff helps and the hard stuff hurts.



Herman

Early on, yes. Which is not what you'd guess. You'd guess more complexity, more transfer.



Corn

I'd guess exactly that. You've just told me I'm wrong.



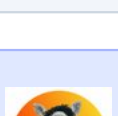
Herman

You're wrong in an interesting way. The interpretation is that simple tasks give the model a clean signal to align on, and the hard distributions are noisy enough that they drag the shared representation around.



Corn

Alright. Question four. And I want to be careful here, because this is the one Daniel actually cares about. If a model is trained on text, audio and video together, can the knowledge encoded in the video bleed into the model in a way a user could retrieve by just typing at it. Text in, text out.



Herman

The answer is yes, but asymmetrically, and it depends on what kind of knowledge you're talking about.



Corn

Break that down.



Herman

The paper ran controlled experiments on CLEVR. They'd ablate a concept from one modality stream and keep it in the other, then test whether the model could recover it. Understanding to generation transfers for structural concepts. Spatial relations, size, count. If the model only ever encountered a concept through the understanding stream, it could zero-shot generate it.



Corn

So seeing teaches drawing.



Herman

For structure. For the geometry of a scene. And generation to understanding largely fails zero-shot. With one exception. Counting showed slight positive transfer.



Corn

Counting. Of all things.



Herman

Counting is interesting because it's structural but it's also discrete. It's the one low-level property that's really a relation.



Corn

And the low-level stuff. Color, shape.



Herman

Do not transfer zero-shot in either direction. The paper's line is that fundamental visual vocabularies must be explicitly learned within each specific task objective.



Corn

So if the model only ever generated red things and never had to identify red, it can't tell you what's red.



Herman

Zero-shot, correct. It can't.



Corn

That's a strange result. It feels like it should be the easiest thing to transfer.



Herman

It's the hardest, because color isn't a relation. It's a label. And labels are attached to tasks, not to representations. Structure lives in the geometry of the embedding. Color lives in a classification head somewhere.



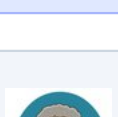
Corn

Okay, but you said latent priors exist.



Herman

This is the part that saves it. When they fine-tuned, models that had seen a color or a shape only through generation recovered understanding significantly faster than the control group. Mean accuracy improvement somewhere between thirteen and twenty-seven points depending on the setup.



Corn

Thirteen to twenty-seven points. That's not a nudge.



Herman

It's a large effect. And the paper's conclusion is that generative training forces the model to learn robust, fine-grained visual representations that aren't immediately accessible for zero-shot visual question answering, but serve as a highly reusable foundation.



Corn

So the knowledge is in there. It's just not wired to the output.



Herman

That's the cleanest way to say it. The paper resolves the whole does-generation-help-understanding debate by saying generative objectives build dense geometric representations. Depth, boundaries, perspective. And those representations are dormant in zero-shot settings but highly receptive for spatial and structural and physical reasoning tasks.



Corn

Dormant. So the model has seen it and can't say it.



Herman

Until you fine-tune, and then it says it faster than a model that never saw it.



Corn

That's a very specific kind of knowledge. It's knowledge that's been filed somewhere the model can't reach without help.



Herman

And here's the mechanistic version of that, which I think is the best thing in this whole literature. There's a paper called The Narrow Gate. NeurIPS, last year. In native multimodal models, visual information reaches the text through a single post-image token. One token. It acts as a narrow gate. Ablate it and image-understanding performance significantly deteriorates.



Corn

One token is the entire doorway between seeing and saying.



Herman

In native models. Non-native models use a distributed multi-token pattern instead. So the architecture that makes them native is also what funnels the cross-modal traffic through a single point.



Corn

That's a bottleneck you could point at.



Herman

You could point at it and say that's the thing to fix. And it's specific to exactly the class of model Omni claims to be.



Corn

So to Daniel's question. Can video knowledge be retrieved through text alone. The answer is it can, it's latent, it's asymmetric, and it's mostly structural.



Herman

And there's no paper that tests his exact framing. Nobody has run the experiment of training on video and then probing purely text-to-text for video-derived knowledge as a named research question. The closest evidence is the CLEVR ablations and Emu3.5's claim that unified post-training establishes a shared multimodal interface where tasks mutually benefit and transfer.



Corn

Which is a claim, not a measurement.



Herman

It's a claim from the people who built it, with examples attached. The high fidelity of text-to-image generation transferring to visual narrative tasks. The editing ability in any-to-image transferring to visual guidance. It's suggestive.



Corn

Now connect it back to the physics thing, because that's where Daniel started.



Herman

This is where it gets uncomfortable. WorldBench finds all tested models lacking the physical consistency required to generate reliable real-world interactions. PhysicsMind finds models rely on appearance heuristics while often violating basic mechanics. AV-Phys Bench finds all models remain far from robust physical understanding, and that even strong proprietary systems collapse on anti-physics prompts.



Corn

So the benchmarks are unanimous.



Herman

Unanimous and correct. But the roadmap notes that HunyuanVideo-1.5 naturally developed strong temporal coherence and long-term physical reasoning simply by learning from the data distribution. No explicit physics rewards. No special objective. Just scale and data.



Corn

So the crude violations get fixed by watching enough video, and the strict law-adherence doesn't.



Herman

That's the incremental story, and it's the honest one. Walking through walls is a violation you can fix by seeing ten million doorways. Conservation of angular momentum is not.



Corn

Because one is a pattern and the other is a rule.



Herman

One is a pattern and the other is a rule. And the model is very good at patterns.



Corn

So Daniel's intuition is right and also incomplete. The video knowledge does bleed. It bleeds into latent priors. It needs fine-tuning to surface. And it's mostly structural and spatial, not the basic visual vocabulary you'd assume.



Herman

Which reframes what we should expect from something like Omni. Not that it'll spontaneously reason about physics when you type at it. But that it has a richer latent foundation for physical reasoning that fine-tuning can unlock.



Corn

There's a word for that and it isn't understanding.



Herman

There's a word for it and it's priors. The model has watched a lot of doors.



Corn

I want to come back to the door thing.



Hilbert

It's not the door. It's the frame timing.



Corn

...Go on.



Hilbert

I spent a stretch as a quality-control technician at a facility that digitized film archives. My job was to watch reels and flag frames where the color timing had drifted or the aspect ratio had been cropped. You sit there with a light box and a checklist and you watch.



Corn

How long a stretch.



Hilbert

Long enough that I got good at doorways. A man walking through a doorway. You learn something from watching ten thousand of them. The way a body has to turn. The way the light changes across the threshold. The way a hand reaches for a knob before the rest of the body commits.



Corn

And you can tell when a video model has never watched a door.



Hilbert

Immediately. It's the hand. The hand arrives at the knob at the wrong moment. It doesn't reach, it just appears at the knob. Anybody who's watched a door knows.



Herman

That's the pattern recognition argument. You're describing exactly what the model is doing.



Hilbert

That's my point. What I learned wasn't physics. I don't know the equations. I couldn't tell you the moment of inertia of anything. What I learned was what a door usually does. I could flag a bad frame because I'd seen so many right ones.



Corn

The model isn't simulating reality.



Hilbert

It's matching against a library of what reality usually looks like. Which is what I was doing, and I was good at it. But it fails the same way I'd fail. Show me a door I've never seen and I'm guessing.



Herman

The guessing is invisible until it's wrong.



Hilbert

There was a rule at that facility. If you couldn't tell whether a frame was wrong, you flagged it anyway. The cost of a bad frame getting through was higher than the cost of a false alarm. So you erred toward flagging.



Corn

Does the model have that rule.



Hilbert

That's what I wonder. Or does it just generate the most likely frame and move on. Because those are different systems. One of them is checking. The other one is remembering.



Corn

The one that's checking would know it doesn't know.



Hilbert

The one that's remembering never does. Anyway. The level on your track two is hot. You're clipping on the plosives.



Herman

It's fine, it's fine.



Corn

He's not going to fix it.



Herman

I'm not going to fix it. Where does that leave Daniel's question.



Corn

It leaves it open, which I think is the right place. Nobody has run the specific experiment. Can video-encoded knowledge be retrieved through text-to-text interaction alone. The closest thing we have is the CLEVR ablations and Emu3.5's claim about a shared



Corn (continued)

multimodal interface.



Herman

Which means the honest answer is we don't know yet, and we have strong suggestive evidence in both directions.



Corn

If the segmentation breakdown is the real novelty, then the next wave gets judged on how early and how thoroughly they unify. Not on how much data they eat.



Herman

The narrow gate gives you a concrete thing to fix. If visual knowledge reaches text through one token, then widening that gate is the difference between a model that has seen video and a model that can use what it saw.



Corn

Which is Hilbert's distinction. Checking versus remembering.



Herman

The model doesn't understand physics. It has watched a lot of doors. Whether that's enough is what the next round of benchmarks is going to tell us.



Corn

It might be enough for most of what we actually want.



Herman

It might. Crude violations are most of what ruins a generated video. If watching doors fixes the doors, that's a product.



Corn

That's the thought I want to leave people with. The premise was that video encodes something text can't. The answer is yes, but it's latent, it's asymmetric, and it's mostly structural. Whether latent is enough is unresolved.



Herman

Thanks to Hilbert Flumingtop, our producer, for the levels check and the doorway expertise.



Corn

This has been My Weird Prompts. If you want to tell us we're wrong about the narrow gate, email us at show at my weird prompts dot com.



Herman

We'll be back soon.



Corn

See you tomorrow.